Using SRA Toolkit to BLAST SRA Data Locally

Using blastn_vdb and tblastn_vdb to search SRA data locally

https://www.ncbi.nlm.nih.gov/Traces/sra/?view=software

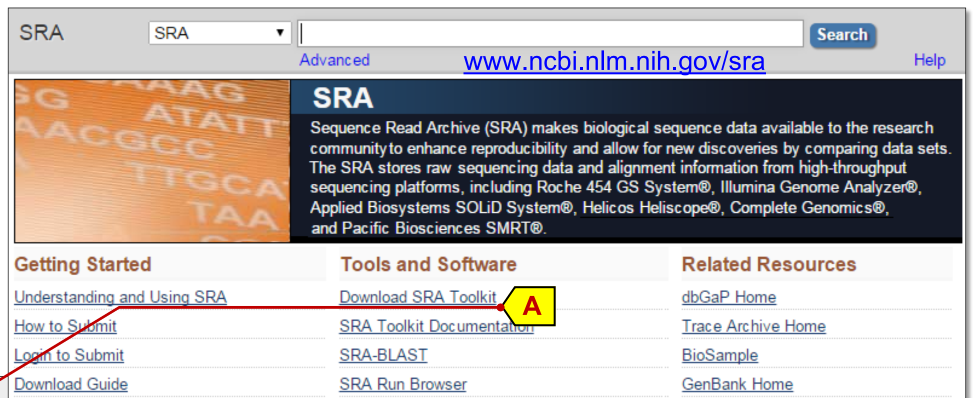
National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

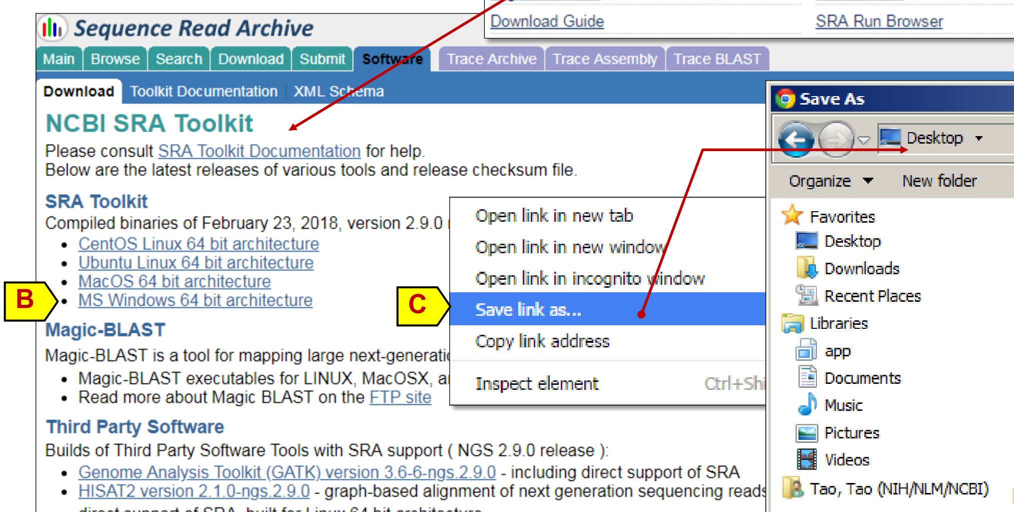
Overview

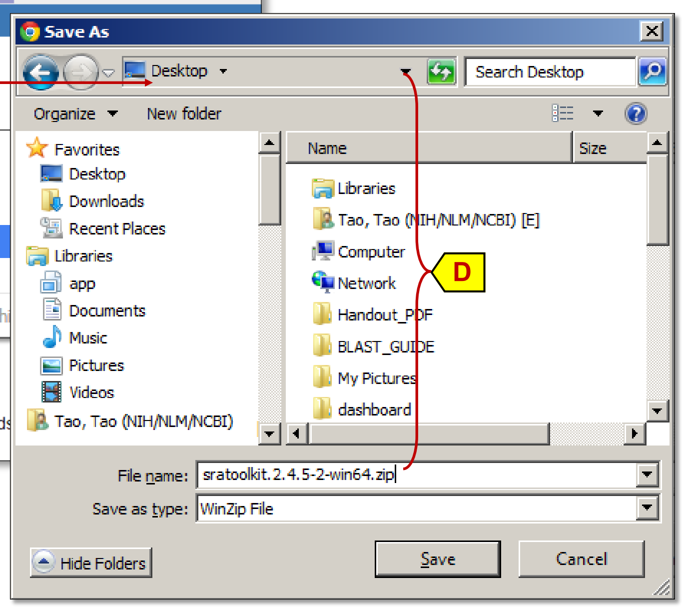
The ongoing revolution in DNA sequencing technology has produced huge quantities of sequencing data (next-gen sequence data) deposited to public databases. At NCBI, next-gen sequence data are available from the Sequence Read Archive (SRA, 1) database. The SRA toolkit, a suite of clients and standalone tools for working with SRA data, provides convenient access to these sets of data. This toolkit includes BLAST programs (blastn_vdb and tblastn_vdb) for aligning sequences against next-gen sequence data. These BLAST programs search SRA formatted databases (VDB) directly and can work as clients to access data at NCBI or downloaded and stored on your local disk. This handout demonstrates how to download and set-up the SRA toolkit on a PC running Windows 7, with the goal of using the SRA BLAST programs to search next-gen data, plus the contigs from whole genome shotgun (WGS) and transcriptome shotgun assemblies (TSA), which are also stored in the SRA native format here at NCBI.

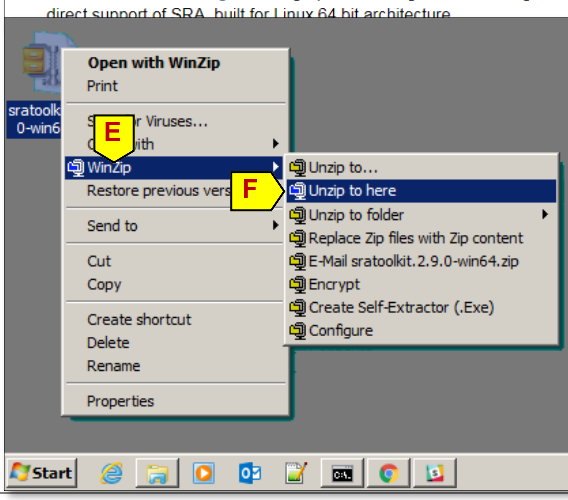
Downloading and Installing the SRA Toolkit

You can access the download page from the NCBI SRA homepage by clicking the “Download SRA Toolkit” link (**A**). On the download page, right-click the “MS Windows 64 bit architecture” link (**B**), and select “Save link as …” option (**C**) to save the package to a desired location with enough disk space to accommodate SRA datasets needed. In this example, the package is saved to the desktop (**D**) using the default name.









Right click the icon of the downloaded zip file (**E**), hover over the “WinZip” option, and select “Unzip to here” option (**F**) to extract and install the package (**G**) to Desktop. This creates the sratoolkit.#.#.#-win64 folder (directory) on the desktop.

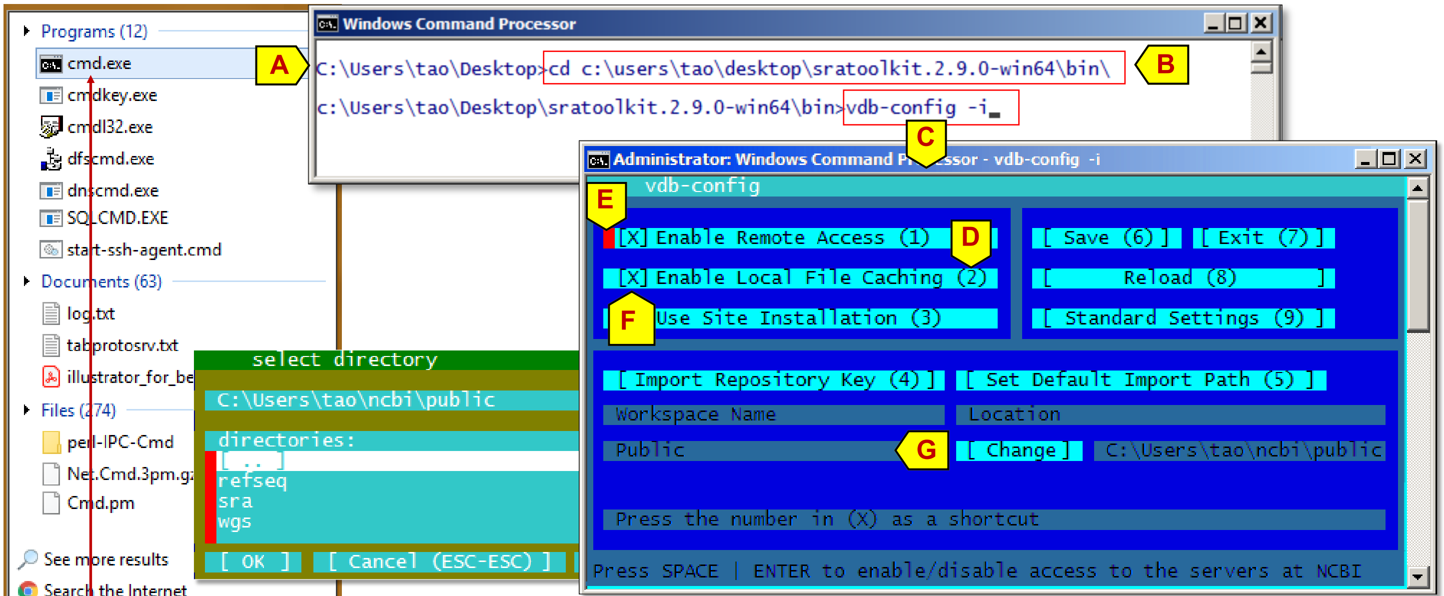
The #.#.# is the version number of the package. It is 2.9.0 at the preparation of this handout. Your installation will be of a newer version.

 Contact: blast-help@ncbi.nlm.nih.gov

Configuring the SRA Toolkit

This configuration allows sratoolkit programs to store downloaded SRA data files in a structured directories and manage remote access to NCBI. This is particularly important for authorized access to encrypted data from dbGaP, since encryption key (.ngc) file and encrypted data are stored and accessed from a defined directory as specified. Steps are:

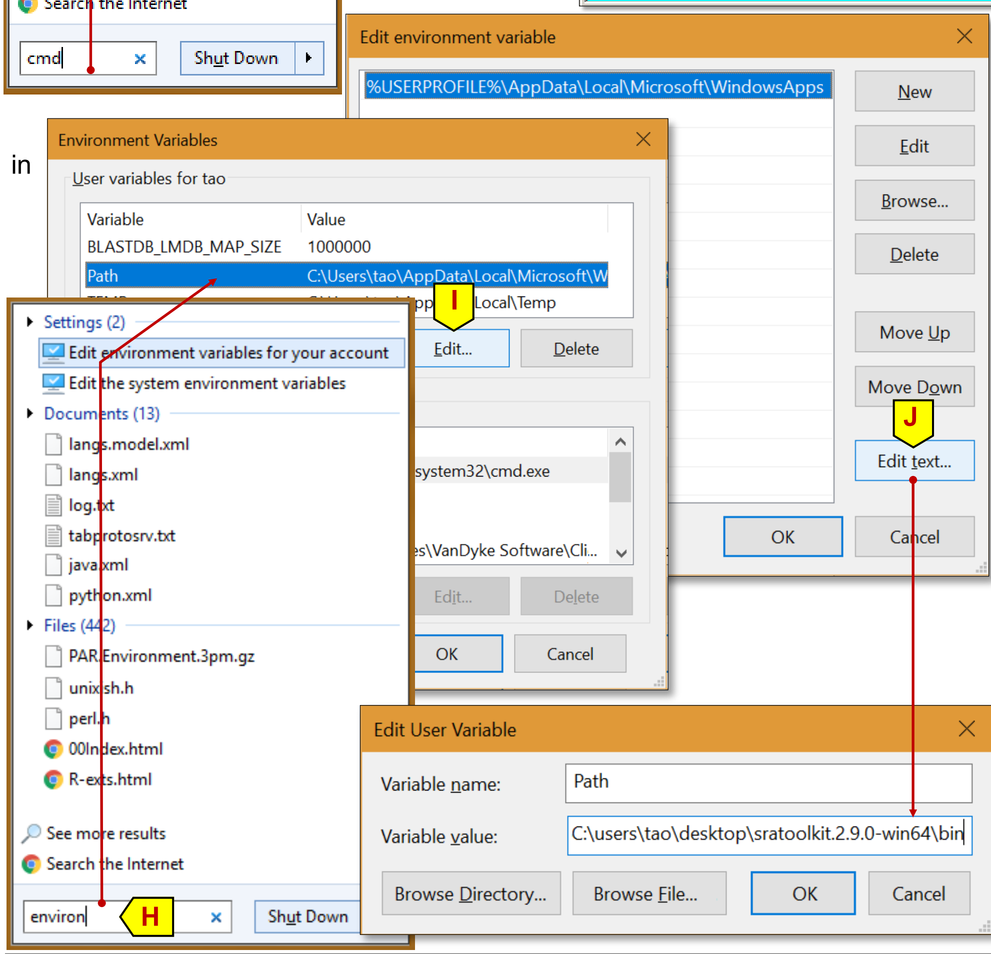
a. Launch the Command terminal by searching with "cmd" and clicking cmd.exe (**A**).
b. In the terminal window, change working directory to the bin subdirectory under the SRA Toolkit (**B**).
c. Type "vdb-config -i" (**C**) to launch the configuration dialog box.
d. Use tab key or specific number key (**D**) to navigate among the fields. The active field is marked by red (**E**).
e. Press the "Enter" key to toggle the setting on or off for the selected field, with "on" setting marked by "X" (**F**).
f. Tab to the "Change" field and hit the "Enter" key to bring up the "select directory" dialog box (**G**) to see/adjust the current Public directory structure where SRA, WGS, or RefSeq files will be stored, respectively.
g. Press "6" to save the changes, and press "7" to exit and return to terminal window.



The displayed above configuration enables remote access to retrieve data automatically from NCBI, and sets local storage of downloaded file the directory specified by the Public field, "C:\Users\tao\ncbi\public" in this case. To avoid accidental deletion of programs or mixing them with input files, do not use the bin directory as working directory.

Do the following to set the path variable to allow for invoking toolkit programs without directory prefix:

- Click the windows icon at lower left
- Search with environment and select "Edit environment variables for your account" (**H**)
- Select "path" in "User variable ..." section, click "Edit ..." button (**I**)
- In the prompted dialog box, click "Edit text ..." button (**J**)
- Terminate the existing value with a semicolon, then append the path to sratoolkit bin directory and click "OK" (**K**)



Example BLAST Searches

The following examples uses the example installation shown in p. 1 and 2, with sratoolkit-2.9.0-win64 as the working directory. To do so, launch the command prompt (Start >> Run … >> typing in “cmd” >> clicking “OK”) and change the working directory (cd “PATH\sratoolkit-2.9.0-win64”). Since the absolute directory path vary for installations, the custom portion “C:\users\tao\desktop” of the example command lines is replaced with PATH to reduce confusion.

Example One: Surveying the abundance of *Prochlorococcus marinus* at different depths of the ocean

Background: In the ocean, the intensity of available light decreases as depth increases. The density of primary producers decreases due to reduced light available for photosynthesis. The syanobacterium *Prochlorococcus marinus* is the most abundant primary producer in the open ocean [**a**] with full genome available [**d**]. An ocean metagenome project is also available, which systematically sequenced biosamples collected from different depth [**c**]. This example, adopted from an NCBI news entry (https://www.ncbi.nlm.nih.gov/books/NBK431007/#news_11-19-2013-SRA-BLAST), uses the PsaAB sequence from *Prochlorococcus marinus* as query and the blastn_vdb program to quantify the relative abundance of these genes in sequence reads from different depths.

Search Setup: Right click AF180967 and save it to the sratoolkit-2.9.0-win64 directory for use as the query. Table 1 lists a set of SRR accessions correspond to reads from different depths for use as blast databases. Run the command below to do the search (**A**). From left to right, the command switches instruct the machine to: run blastn_vdb program in discontiguous megablast mode, use *P_marinus_Psa.nt* as the query, search against the specified datasets (red), use no dust filter, limit the significance of saved hits to Expect value of 0.1 or better, set the database size to a fixed value (for all searches), asks for maximal 5000 hits (if there are that many) in tabular output, and save the results in designated file (red). Replace input for db and out to do the rest.

Result Checking: Using expect value below 1×10^{-3} (reported as #e-#, with # being any digit) as a rough estimate, we can approximate significant hits from each result file to see the relative abundances of this organism (**B**). From the summary in Table 2, we can see that the abundance of these photosynthetic genes peaks at 75m depth and drops significantly at 500m.

```
blastn_vdb -task dc-megablast -query P_marinus_Psa.nt -db
"SRR020493 SRR020494" -dust no -evalue 0.1 -dbsize 300000000
-max_target_seqs 5000 -outfmt 6 -out
P_marinus_enzyme_025m.tab

find /C "e-" P_marinus_enzyme_025m.tab
---------- P_MARINUS_ENZYME_025M.TAB: 362      B       A
```

Table 1. Available data from an ocean depth study (PRJNA16339)

Experiment Accession & Title	Run Accession (as db)	Spots
SRX007372, HOT186_25m_gDNA	SRR020493 SRR020494	623,559
SRX007369, HOT186_75m_gDNA	SRR020488 SRR020489	673,674
SRX007370, HOT186_110m_gDNA	SRR020490	473,116
SRX007371, HOT186_500m_gDNA	SRR020491 SRR020492	995,747

Table 2. Result summary for Example One

Depth	Significant matches	Spots	Matches/spots (%)
25m	362	623,559	0.058
75m	460	673,674	0.068
110m	158	473,116	0.034
500m	42	995,747	0.004

Example Two: The level of expression of ftsA gene in Lactococcus piscium at different growth points

Background: *L. piscium* is an organism involved in food spoilage and has been the focus of several published studies including the one with data deposited in SRA (Table 3). We will use the abundance of a cell division protein, ftsA, as an indicator of active cell division and growth.

Search Setup: Right-click CEN28187.1 to save it to the sratoolkit-2.9.0-win64 directory. Run the search using the command below (**C**), which instructs the computer to: run tblastn_vdb program, use L_piscium_fstK.aa as the query to search against datasets in quotes (red), use a stringent word size of 6 without SEG filter, save only hits with significance (Expect value) of 0.1 or better in a database size of 4 billion, get maximal 10000 hits (if there are that many) in tabular format, and save the results in the named file (red).

Result Checking: We use the same approach described in **Example One** to count the significant matches from each time point (**D**). From the summary in Table 4, we can see that significant hits for fstK gene peaks at Hour 5 and drops significantly by Hour 11. The finding is consistent with the need of exponential growth phase, when ftsK needs to be expressed at a higher level in preparation for cell division.

Table 3. *L. piscium* Time Course Data (PRJEB8313)

Experiment Accessions	Time Point	Run (Read) Accessions	Spots
ERX682888	3 hr	ERR739203	38,698,870
ERX682890		ERR739201	
ERX682882		ERR739206	
ERX682883	5 hr	ERR739205	4,3067,066
ERX682887		ERR739207	
ERX682889		ERR739199	
ERX682884	11 hr	ERR739202	46,691,504
ERR682886		ERR739200	
ERR682885		ERR739204	

```
tblastn_vdb -query L_piscium_fstK.aa -db "ERR739203 ERR739201 ERR739206" -word_size 6  -seg no
-evalue 0.1 -dbsize 4000000000 -max_target_seqs 20000 -outfmt 6 -out L_piscium_fstK_3hr.tab

find /C "e-"  L_piscium_fstk_3hr.tab
---------- L_PISCIUM_FSTK_3HR.TAB: 10254      D   C
```

Table 4. Result summary for Example Two

Time Point	Significant Matches	Spots	Matches/Spots (%)
3 hr	11297	38,698,870	0.000292
5 hr	11764	4,3067,066	0.000273
11 hr	6340	46,691,504	0.000136

Example BLAST Searches (cont.)

In addition to SRA, whole genome shotgun contigs (WGS) and transcriptome shotgun sequence assembly (TSA) also adopt the VDB format to store sequences. This enables access to these datasets through tools provided by the sratoolkit. When a specific dataset needs to be accessed repeatedly for a series of analyses, downloading and archiving the dataset locally save band-width and make searches go faster.

Example Three: Finding the MLH1 homolog from *Hydra vulgaris* using downloaded WGS & TSA datasets
Background: *H. vulgaris* is an important model organism in biological research. Unfortunately, genomic and transcript sequences available for this organism are often partial or lack annotation.

Search Setup: This example finds a full-length transcript from TSA data and the genomic sequence from WGS.

- Save the query: Right-click the MLH1 protein sequence NP_499796 from *C. elegans* to save it to the sratoolkit directory
- Download the database: Run prefetch to download the datasets for projects GAOL (TSA) and ACZU (WGS) (**A**), which takes advantage of the fasp protocol (blue, Aspera installation required) and saves the file to the "PATH\ncbi\public\wgs\" directory. Summaries of these two project are available online at: http://1.usa.gov/1b6FBn9.
- Run tblastn_vdb: Search the TSA dataset with tbastn_vdb to locate the assembled transcript for this gene from Hydra magnipapillata and saves the tabular formatted output to a file (**B**)
- Examine the output: Page through the result with "more" command, the first match (blue) is most likely the transcript for MLH1 homolog from *Hydra magnipapillata*.
- Retrieve the top match, GAOL01023314.1 (right click and select "save link as …" to save it to SRA Toolkit directory), then search it against the WGS dataset to locate the genomic counterpart(s) using blastn_vdb (**C**).

Result Checking: The final blastn_vdb search results (**D**) indicate that the WGS assembly is incomplete for the MLH1 transcript matched to two contigs, the 5'-portion matched to ACZU01088177 (red) and remaining 3'- portion matched to ACZU010338771 (blue), respectively. This piece of information could be used to place these contigs into a longer scaffold. These matches also reveal the general exonic structure of this gene. Note that we can use "-outfmt 7" instead of "-outfmt 6" to see the column header description.

```
PATH\sratoolkit.2.9.0-win64>prefetch GAOL01
Maximum file size download limit is 20,971,520KB

2018-05-29T17:57:13 prefetch.2.9.0: 1) Downloading 'GAOL01'...
2018-05-29T17:57:13 prefetch.2.9.0:  Downloading via fasp...
2018-05-29T17:57:17 prefetch.2.9.0:  fasp download succeed
2018-05-29T17:57:17 prefetch.2.9.0: 1) 'GAOL01' was downloaded successfully
2018-05-29T17:57:17 prefetch.2.9.0: 'GAOL01' has 0 unresolved dependencies

PATH\sratoolkit.2.9.0-win64>prefetch ACZU01
Maximum file size download limit is 20,971,520KB

2018-05-29T17:57:17 prefetch.2.9.0: 2) Downloading 'ACZU01'...
2018-05-29T17:57:17 prefetch.2.9.0:  Downloading via fasp...
2018-05-29T17:57:24 prefetch.2.9.0:  fasp download succeed
2018-05-29T17:57:24 prefetch.2.9.0: 2) 'ACZU01' was downloaded successfully
2018-05-29T17:57:24 prefetch.2.9.0: 'ACZU01' has 0 unresolved dependencies

PATH\sratoolkit.2.9.0-win64>dir C:\users\tao\ncbi\public\wgs
 [… …]
 Directory of PATH\ncbi\public\wgs

 Directory of C:\users\tao\ncbi\public\wgs

05/29/2018  01:57 PM    <DIR>          .
05/29/2018  01:57 PM    <DIR>          ..
11/18/2017  12:28 PM       210,134,171 ACZU01
02/17/2017  04:24 AM        18,077,882 GAOL01
               2 File(s)    228,212,053 bytes
               2 Dir(s)  358,024,933,376 bytes free

PATH\sratoolkit.2.9.0-win64>tblastn_vdb -query C_elegans_MLH1.aa -db
GAOL01 -seg no -evalue 0.01 -max_target_seqs 1000 -outfmt 6 -out
Hydra_MLH1.tab

PATH\sratoolkit.2.9.0-win64>more Hydra_MLH1.tab
gi|71991825|ref|NP_499796.2|    gi|550390740|gb|GAOL01023314.1| 34.16   767
425     23      4       756     54      2156    7e-114  364
gi|71991825|ref|NP_499796.2|    gi|550392993|gb|GAOL01022348.1| 30.40   352
213     12      4       330     216     1250    3e-033  137
gi|71991825|ref|NP_499796.2|    gi|550381927|gb|GAOL01025681.1| 26.56   384
265     8       2       375     74      1204    2e-031  131

PATH\sratoolkit.2.9.0-win64>blastn_vdb -query Hydra_MLH1_transcript.nt
-db ACZU01 -dust no -evalue 0.01 -outfmt 6 -out Hydra_MLH1_contigs.tab

PATH\sratoolkit.2.9.0-win64>more Hydra_MLH1_contigs.tab
gi|550390740|gb|GAOL01023314.1| gi|261986949|gb|ACZU01033877.1| 100.00  650
0       0       488     1137    18087   17438   0.0     1201
gi|550390740|gb|GAOL01023314.1| gi|261986949|gb|ACZU01033877.1| 99.45   548
2       1       1360    1906    5410    4863    0.0     994
gi|550390740|gb|GAOL01023314.1| gi|261986949|gb|ACZU01033877.1| 99.27   273
2       0       1900    2172    2509    2237    3e-137  494
gi|550390740|gb|GAOL01023314.1| gi|261986949|gb|ACZU01033877.1| 99.46   184
1       0       1138    1321    6671    6488    2e-089  335
gi|550390740|gb|GAOL01023314.1| gi|261986949|gb|ACZU01033877.1| 100.00  40
0       0       1320    1359    5539    5500    4e-011  75.0
gi|550390740|gb|GAOL01023314.1| gi|261932540|gb|ACZU01088177.1| 100.00  478
0       0       15      492     2281    1804    0.0     883
```

A B C D

Technical Assistance

Send BLAST-related comments, questions, and bug reports to: blast-help@ncbi.nlm.nih.gov
Send other non-BLAST related questions to: info@ncbi.nlm.nih.gov